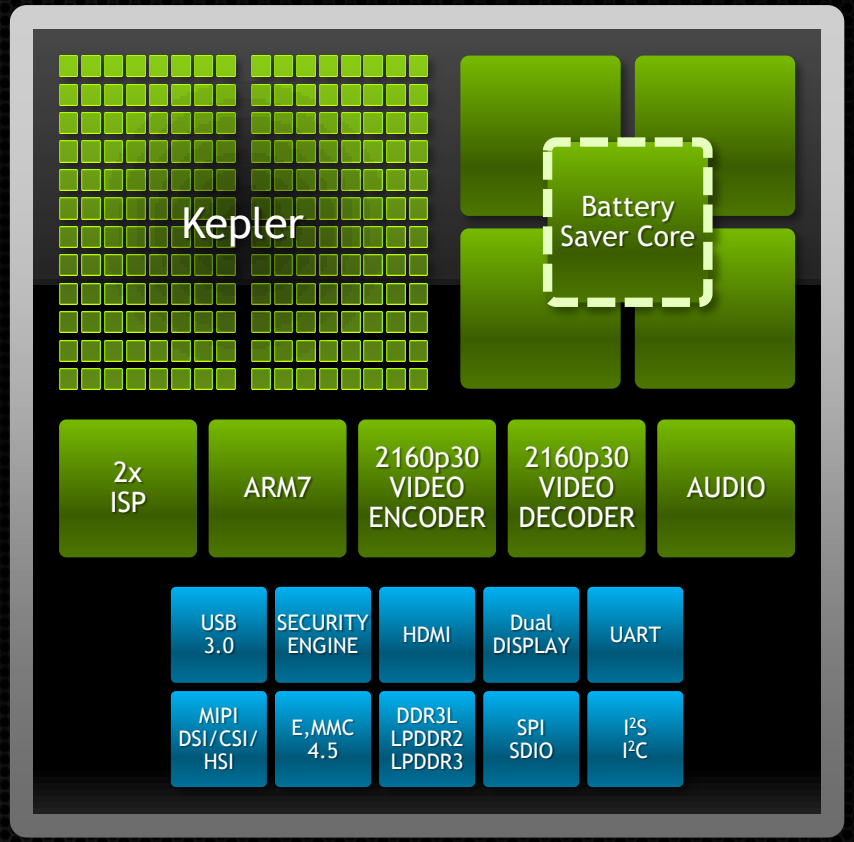
NVIDIA'S TEGRA K1 SYSTEM-ON-CHIP

Michael Ditty, Tegra Architecture

Co-authors: John Montrym, Craig Wittenbrink

Tegra K1

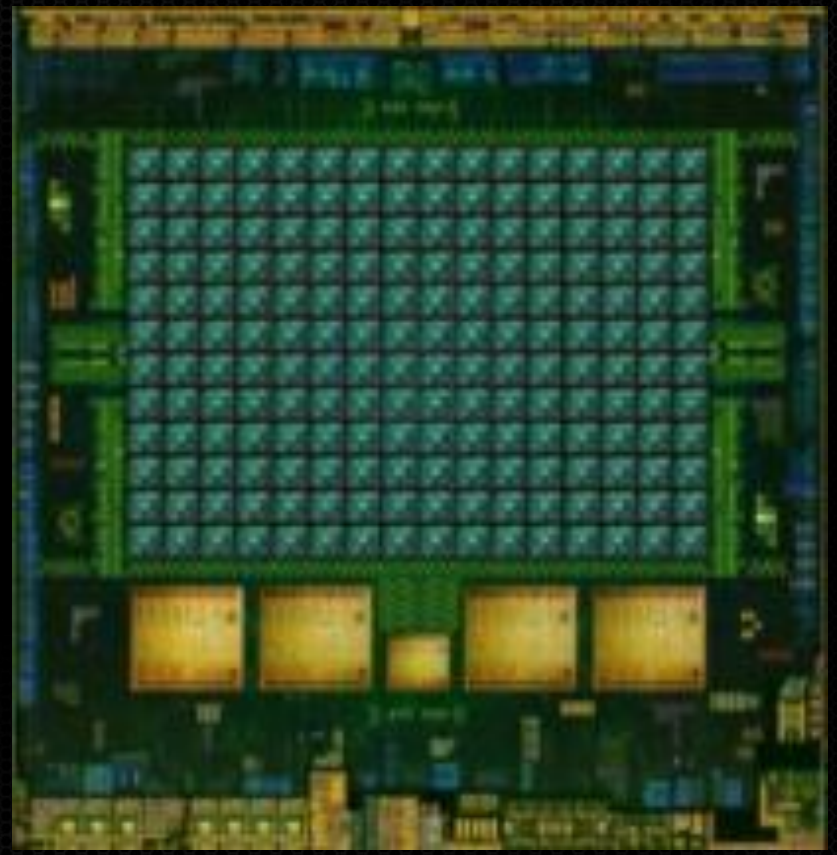


GPU	**Kepler GPU (192 CUDA Cores)** *Open GL 4.4, OpenGL ES3.1+AEP, DX12, CUDA 6*
CPU	**Quad Core Cortex A15 "r3"** *With 5th Battery-Saver Core; 2MB L2 cache* **OR** **Dual Denver CPU**
CAMERA	**Dual High Performance ISP** *1.2 Gigapixel throughput, 100MP sensor*
POWER	**Lower Power** *28HPM, Battery Saver Core*
DISPLAY	**4K panel, 4K HDMI** *DSI, eDP, LVDS, High Speed HDMI 1.4a*

NVIDIA

Overview

- Kepler into Mobile
- Tegra ISP
- Power Management
- Mobile Enablement
- Demo

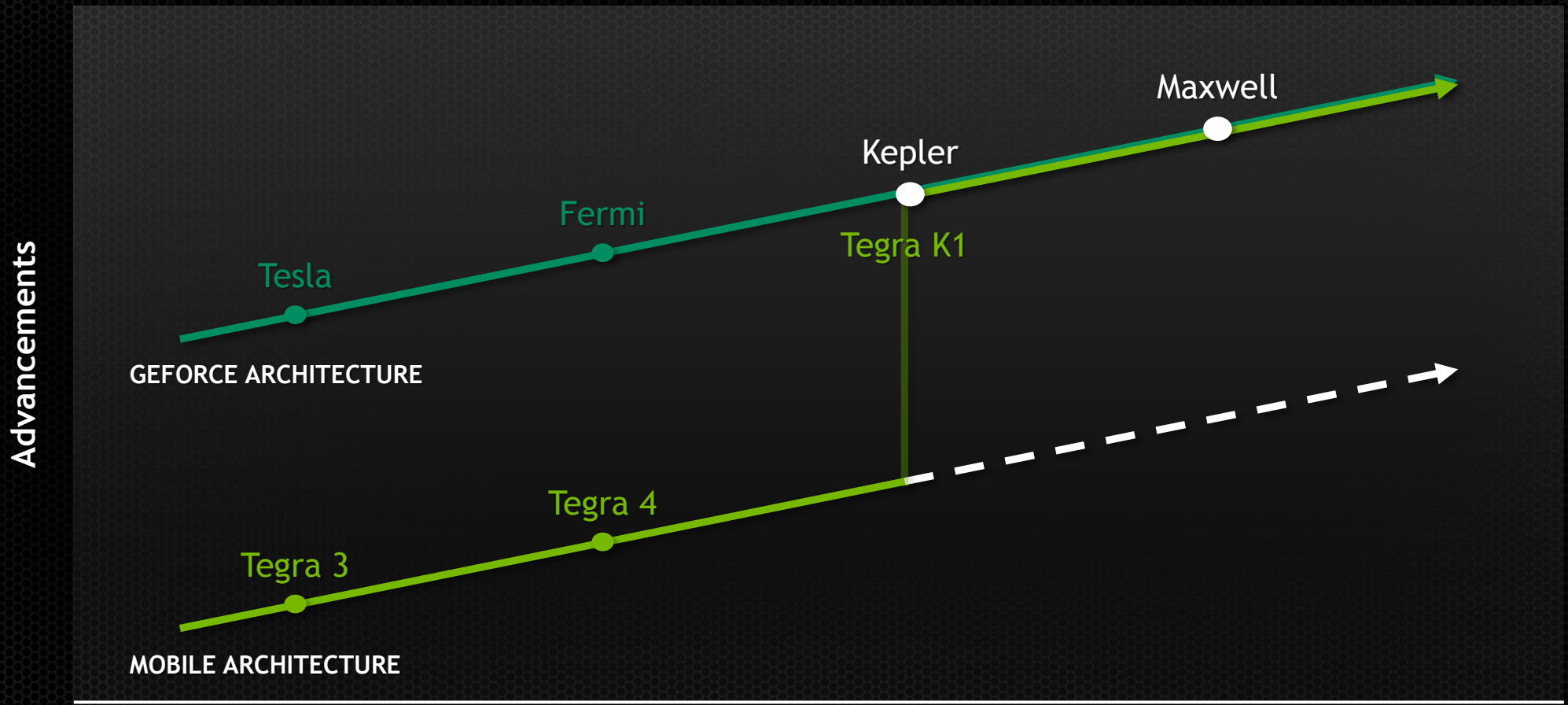


A Major Discontinuity in Mobile Graphics

ES2.0, DX9
Programmable Pixel Shaders

ES3.1+AEP, OGL4.4, DX12
Tessellation, Compute Shaders, ASTC, GPGPU

Mobile Roadmap Meets GeForce



Tegra K1

Metric	Tegra 4	Tegra K1	Units
FP32 ops	48	384	Per clock
Z-only Primitives	0.1	1	Per clock
Zcull	-	256	Pixels/clk
Raster	8	64	Samples/clk
Texture	4	8	Bilinear filters/clk
ZROP	8	64	Samples/clk
L2 size	32	128	KBytes

Tegra K1 / Kepler Graphics Core Architecture

- 192 CUDA cores
- Unified Memory Cache
- Dedicated Accelerators
  - Geom / Tessellation
  - Z Cull
  - Z / Color ROP

NVIDIA

Power Efficiency

- Clock and power gating
  - Multi-level Clock Gating
  - Power Gating
  - Rail Gating
- Architectural power improvements
  - Interconnect and Data Paths architected for mobile
  - Shader Bypass
  - GPU L2 Cache and Compression
- Work reduction
  - Aggressive Culling Of Z, Stencil, Attribute Fetch
  - Early Z

Dual Next Gen ISP

Performance

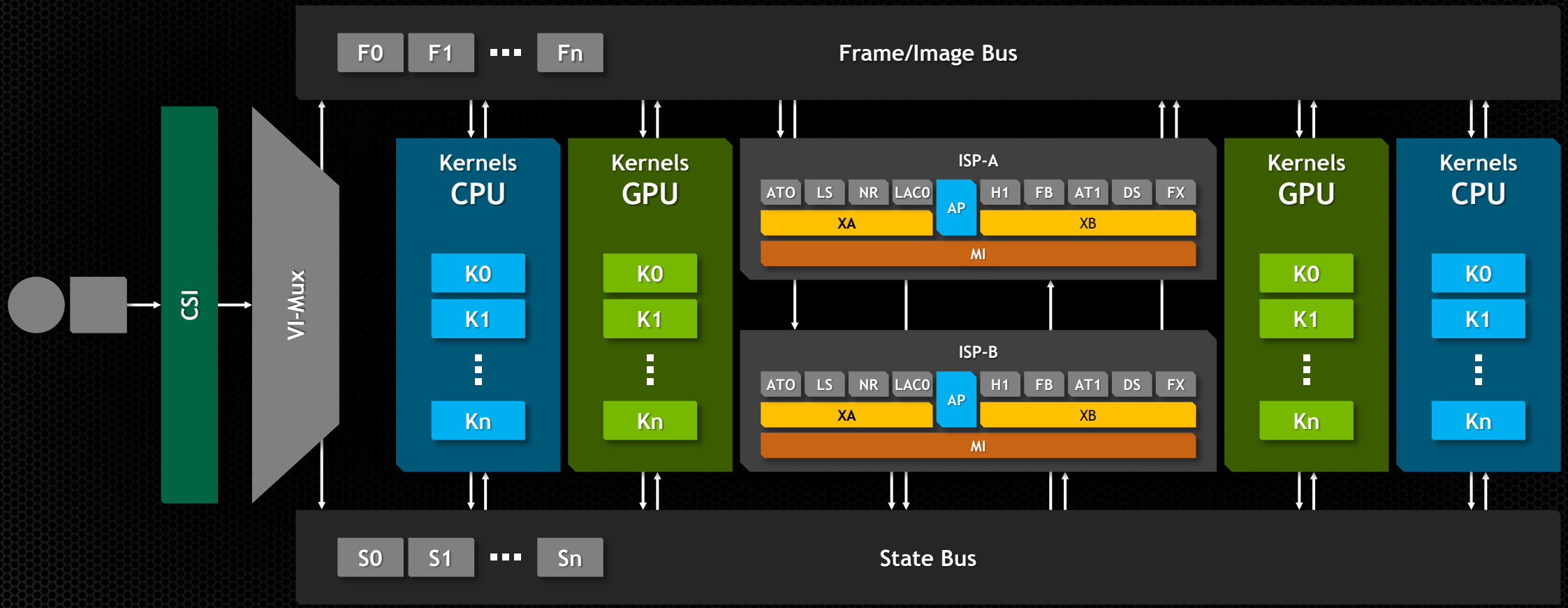
- 1.2Gp total pixel throughput
- 600Mp each ISP
- 4096 simultaneous focus points
- 14 bits input
- 100Mp camera support

Interoperability

- Reconfigurable ISP fabric
- Full GPGPU interoperability
- Memory or Isochronous sourcing

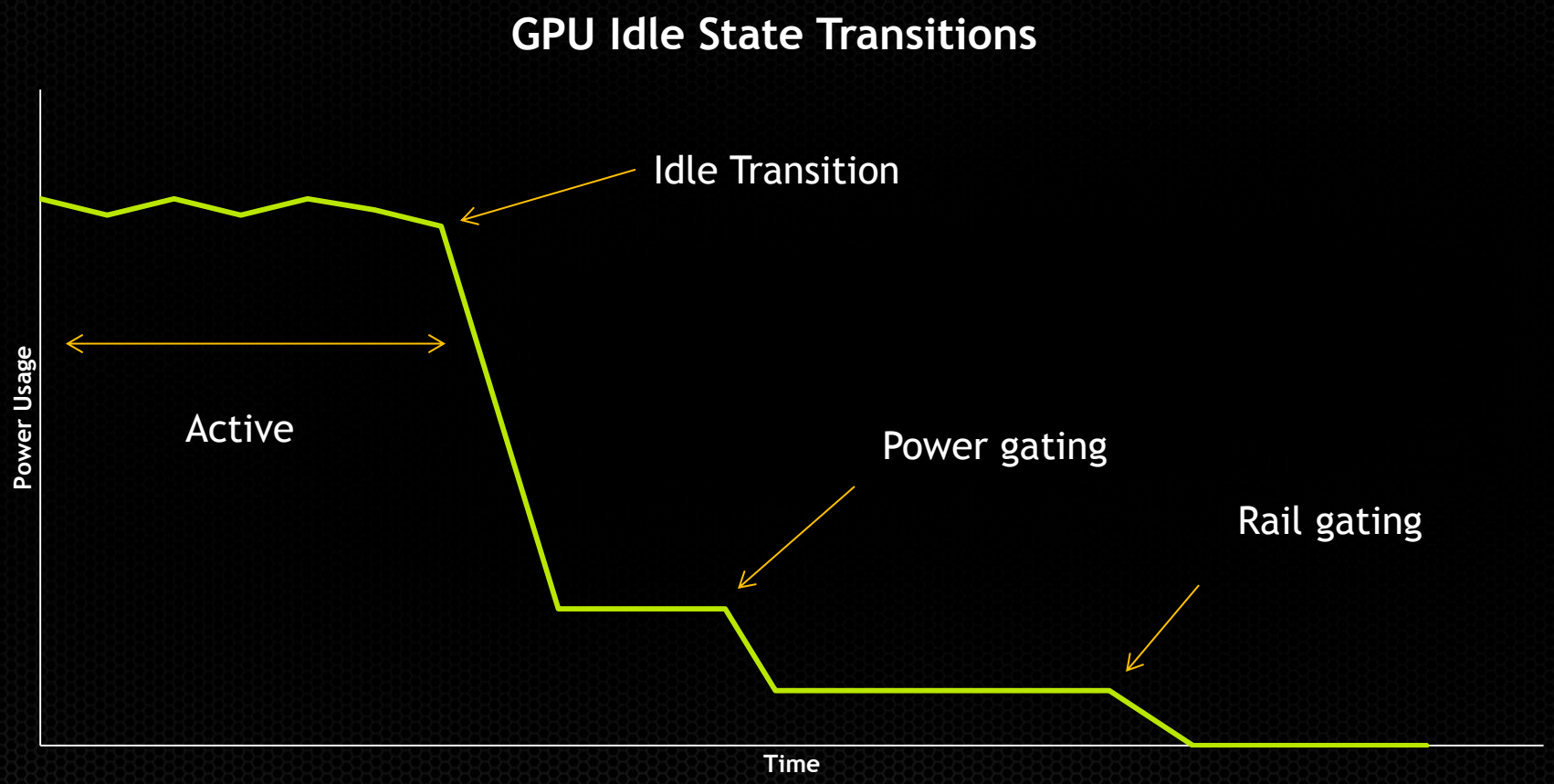
Tegra K1 Computational Photography Architecture

GPU + ISP + CPU



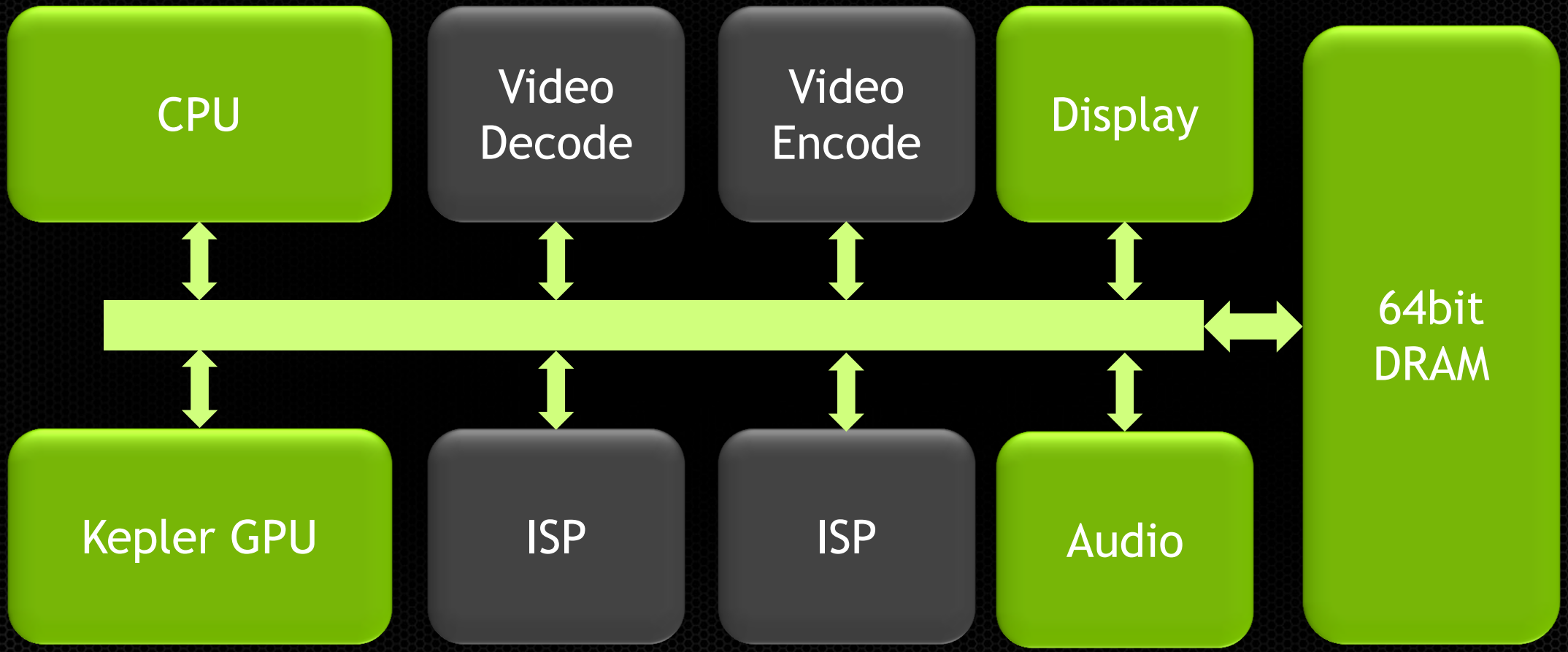
NVIDIA

GPU Power Management



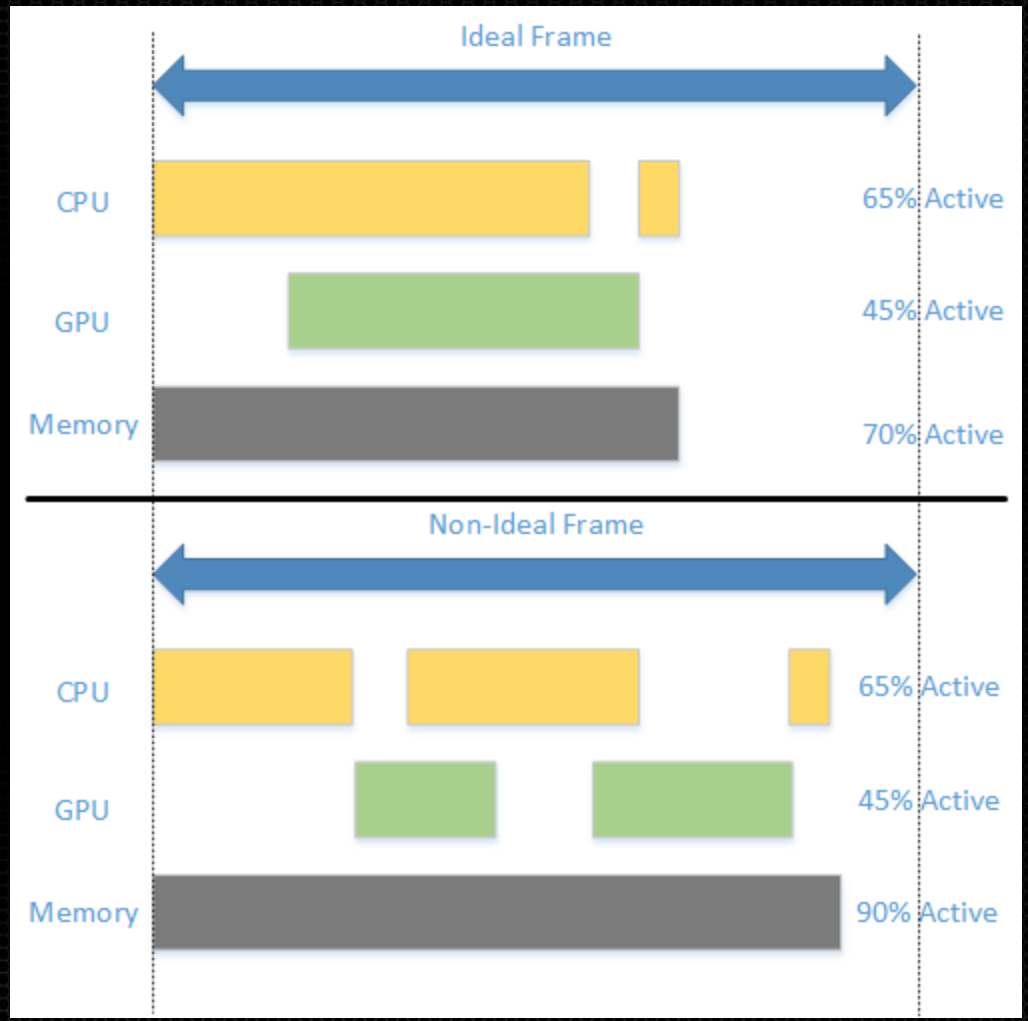
Multi-core Gaming

Multi-core Gaming



Multi-core gaming power management

- Balance power & performance across cores and power rails.
- Clocking policies must look at more than active time.
- Power optimization must be done globally, not locally to each unit.



Multi-core video processing

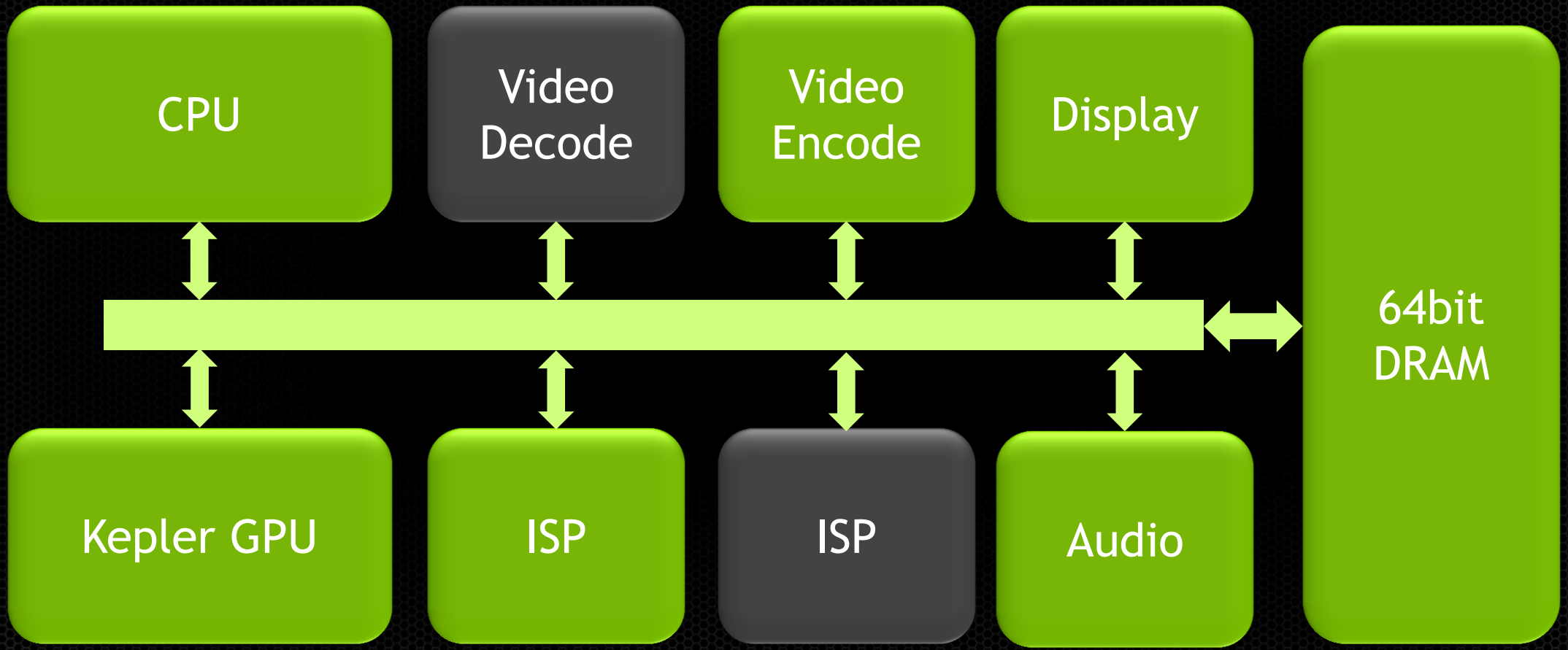
“Live” Local Tone Mapping

Original

Kepler GPGPU Processing 30fps

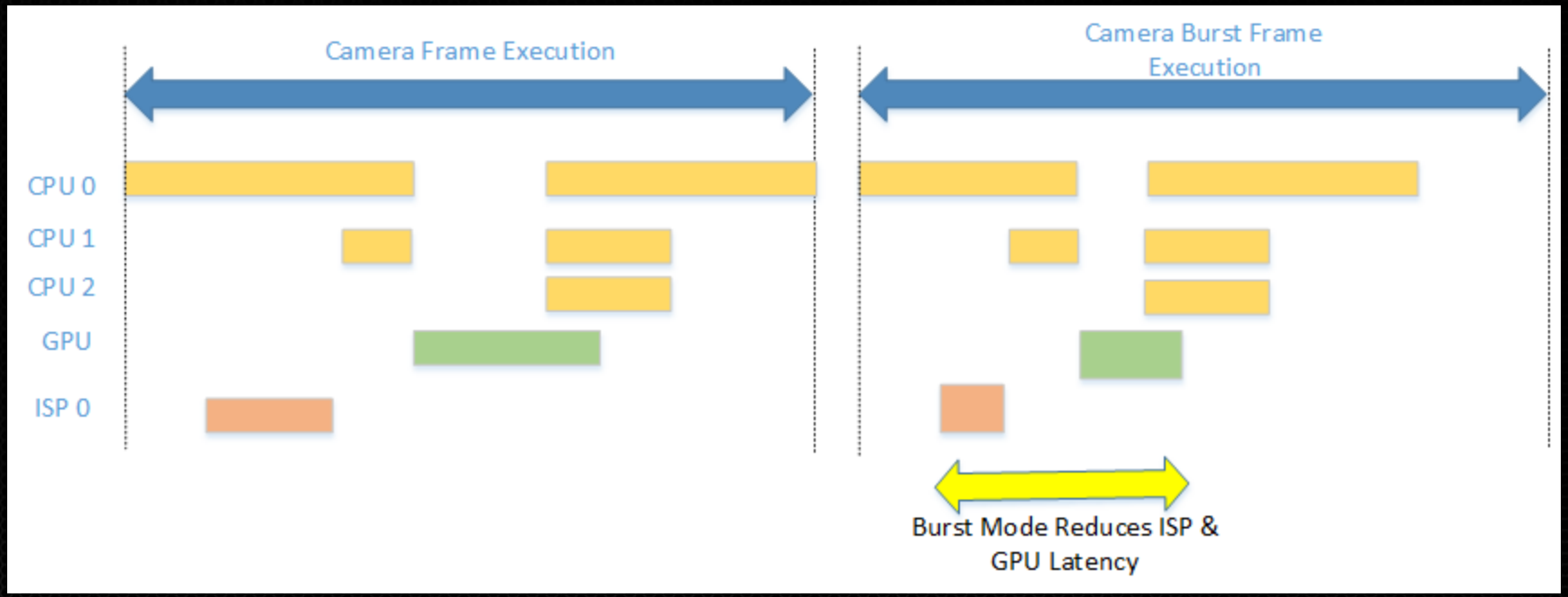
LTM

Multi-core video processing

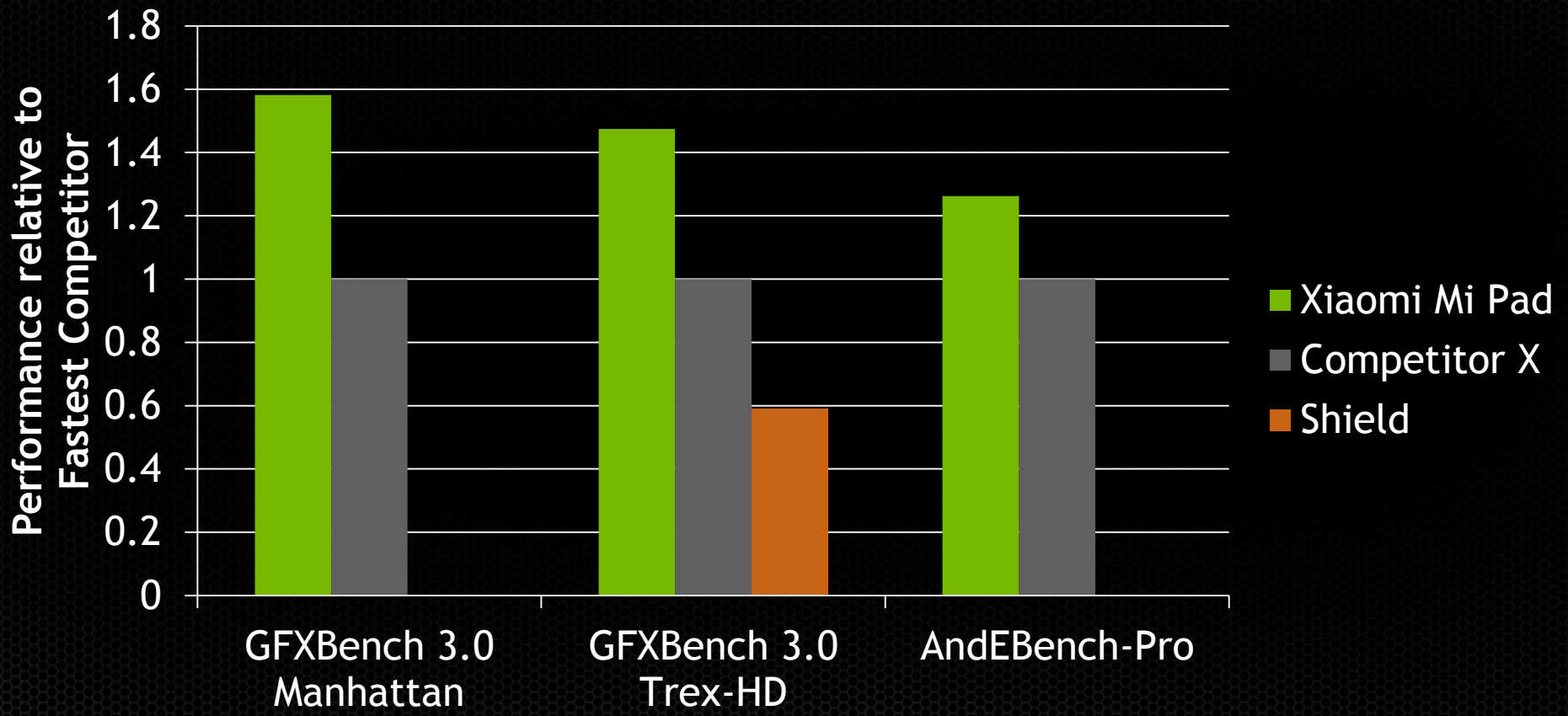


Multi-core video processing

Utilize burst performance for latency reduction



Tegra K1 Benchmarks



NVIDIA

Scalability Across Platforms

Mobile Compute

NV JETSON

- CUDA
- VisionWorks Toolkit
- Renderscript

Tango Tablet

Automotive
Computer Vision

NVIDIA

Consumer Devices

XIAOMI MIPAD

Powered by Tegra K1

Demo

Conclusion

- New capabilities in mobile
  - Compute, OpenGL 4.4, Advanced Imaging Pipeline

- Great performance
  - Over 2x the performance of current mobile devices

- Enabling new platforms and ecosystems

Acknowledgment

We would like to thank the GPU & Tegra teams across NVIDIA who collaborated to make this chip possible.